Maximum A Posteriori Adaptation for Many-to-One Eigenvoice Conversion

Daisuke Tani, Tomoki Toda, Yamato Ohtani, Hiroshi Saruwatari, Kiyohiro Shikano

Graduate School of Information Science, Nara Institute of Science and Technology, Japan
tomoki@is.naist.jp, yamato-o@is.naist.jp

Abstract

Many-to-one eigenvoice conversion (EVC) allows the conversion from an arbitrary speaker’s voice into the pre-determined target speaker’s voice. In this method, a canonical eigenvoice Gaussian mixture model is effectively adapted to any source speaker using only a few utterances as the adaptation data. In this paper, we propose a many-to-one EVC based on maximum a posteriori (MAP) adaptation for further improving the robustness of the adaptation process to the amount of adaptation data. Results of objective and subjective evaluations demonstrate that the proposed method is the most effective among the other conventional many-to-one VC methods when using any amount of adaptation data (e.g., from 300 ms to 16 utterances).

Index Terms: speech synthesis, voice conversion, GMM, eigenvoice, MAP.

1. Introduction

Voice conversion (VC) is a technique for modifying non-linguistic information such as voice characteristics while keeping linguistic information unchanged. One typical example of VC applications is speaker conversion in which a certain speaker’s voice is converted into another speaker’s voice [1]. This technique hopefully provide flexible speech communication over our physical constraints, e.g., speaking in the desired famous character’s voice or generating the own voice of various languages.

As one of statistical VC approaches, the conversion method based on a Gaussian mixture model (GMM) [2] has been well studied. The GMM representing joint probability density of source and target acoustics [3] is trained in advance basically using a parallel data set consisting of utterance pairs of the source and target speakers. Although state-of-the-art conversion methods (e.g., [4]) have dramatically improved the conversion performance, it is still doubtful whether this framework requiring parallel data is acceptable for real users. In order to realize handy VC applications, e.g., the VC system enabling arbitrary users to convert their own voices into the pre-defined target speaker’s voice even if they don’t provide any training data is essential to make the VC framework more flexible.

Recently we have proposed a novel VC framework, many-to-one VC, for realizing the conversion from an arbitrary speaker’s voice into the specific target speaker’s voice [5]. The many-to-one VC framework consists of two main processes, i.e., 1) training and 2) adaptation and conversion. In the training process, multiple parallel data sets consisting of utterance pairs of many pre-stored source speakers and the target speaker are employed for training a canonical conversion model. In the adaptation and conversion process, the canonical conversion model is adapted to arbitrary source speakers using only their speech data without any linguistic restrictions. And then, VC is performed with the adapted conversion model.

Eigenvoice conversion (EVC) [5], which is one of our previously proposed many-to-one VC algorithms [6], enables the unsupervised adaptation process working reasonably well even if using only a few utterances as the adaptation data. This method employs an eigenvoice technique [7] for the GMM-based VC. Because EVC is well described mathematically, it is straightforward to apply well-known techniques such as speaker adaptive training (SAT) [8] for further improving the conversion performance [6]. However, there still remains a problem. When the amount of adaptation data is very limited, i.e., less than one utterance, the conversion performance rapidly degrades due to the over-fitting problem. In order to develop a high-quality VC system to instantly convert input arbitrary speakers’ voices, it is essential to cope with this problem.

In this paper, we apply maximum a posteriori (MAP) estimation [9] to many-to-one EVC for improving the adaptation performance of EVC particularly when the amount of adaptation data is very limited. The prior probability distribution of adaptation parameters for the canonical conversion model is trained through SAT, and then it is effectively used in the adaptation process for robustly estimating the adaptation parameters. We conduct objective and subjective evaluations of the proposed method. Experimental results demonstrate that the proposed method outperforms the other many-to-one VC algorithms with any amount of adaptation data.

The paper is organized as follows. In Section 2, we describe many-to-one EVC. In Section 3, we describe the proposed MAP adaptation for many-to-one EVC. In Section 4, we describe an experimental evaluation. Finally, we summarize this paper in Section 5.

2. Many-to-One EVC

2.1. Eigenvoice GMM (EV-GMM)

We employ 2D-dimensional acoustic feature vectors $X_i = [x_i^T, \Delta x_i^T]^T$ (source speaker’s) and $Y_i = [y_i^T, \Delta y_i^T]^T$ (target speaker’s) consisting of 2-Dimensional static and dynamic feature vectors, where $T$ denotes transposition of the vector. The joint probability density of the source and target feature vectors is modeled by an EV-GMM as follows:

$$P(X, Y | X^{(E)}, \omega) = \sum_{m=1}^{M} \alpha_m N(X, Y;\mu_m^{(X,Y)}(\omega), \Sigma_m^{(X,Y)})$$

where

$$\mu_m^{(X,Y)}(\omega) = \begin{bmatrix} \mu_m^{(X)}(\omega) \\ \mu_m^{(Y)}(\omega) \end{bmatrix} = \begin{bmatrix} B_m^{(X)}w + B_m^{(X)}(0) \\ B_m^{(Y)}(0) \end{bmatrix}$$

and

$$\Sigma_m^{(X,Y)} = \begin{bmatrix} \Sigma_m^{(XX)} & \Sigma_m^{(XY)} \\ \Sigma_m^{(YX)} & \Sigma_m^{(YY)} \end{bmatrix}$$

and $N(\cdot; \mu, \Sigma)$ shows the normal distribution with a mean vector $\mu$ and a covariance matrix $\Sigma$. The weight of the $m^{th}$ mix-
ture component is \( \alpha_m \). The total number of mixture components is \( M \). In many-to-one EVC, the source mean vector of the \( m \)-th mixture component is represented as a linear combination of a bias vector \( b^{(X)}_m(0) \) and representative vectors \( B^{(X)}_m = [b^{(X)}_m(1), \ldots, b^{(X)}_m(J)] \), where the number of representative vectors is \( J \). The \( J \)-dimensional weight vector \( w = [w(1), \ldots, w(J)]^T \) is adapted for individual source speakers while the parameter set of the EV-GMM \( \lambda^{(EV)} \) is tied over different source speakers. This paper employs diagonal covariance matrices for the individual block covariance matrices in Eq. (3).

2.2. Training of EV-GMM

Many-to-one EVC trains the tied parameter set of the EV-GMM in advance using the multiple parallel data sets consisting of many pre-stored source speakers and a single target speaker. We employ SAT [8] to construct a canonical model causing significant improvements of the model adaptation performance. Let \( X_t^{(s)} \) and \( Y_t \) be the feature vector of the \( s \)-th source speaker and that of the target speaker at frame \( t \). SAT estimates not only the tied parameter set \( \lambda^{(EV)} \) but also a set of the weight vectors \( w_t^S = \{w_1, \ldots, w_S\} \) adapted for individual pre-stored source speakers as follows:

\[
\lambda^{(EV)}, w_t^S = \arg\max_{\lambda^{(EV)}, w_t^S} \prod_{s=1}^{S} \prod_{t=1}^{T} P(X_t^{(s)}, Y_t | \lambda^{(EV)}, w_s). \tag{4}
\]

Because the probability density is modeled with a GMM, the following auxiliary function is iteratively maximized with EM algorithm,

\[
Q(\lambda^{(EV)}, w_t^S; \hat{\lambda}^{(EV)}, \hat{w}_t^S) = \sum_{s=1}^{S} \sum_{t=1}^{T} \gamma_{m,t}^{(s)} \log P(X_t^{(s)}, Y_t | m^{(EV)}, w_s). \tag{5}
\]

where

\[
\gamma_{m,t}^{(s)} = P(m | X_t^{(s)}, \lambda^{(EV)}, w_s). \tag{6}
\]

The ML estimates of individual parameters are shown in [6]. The initial model parameters are determined based on principal component analysis (PCA) as shown in [5].

2.3. Adaptation of EV-GMM

The EV-GMM is adapted for an arbitrary source speaker by estimating the optimum weight vector for given speech samples in a completely unsupervised manner, i.e., using neither parallel data nor linguistic information. The weight vector is estimated so that a likelihood of the marginal distribution for a time sequence of the given source feature vectors \( X_1^{(s)}, \ldots, X_T^{(s)} \) is maximized [5] as follows:

\[
\hat{w} = \arg\max_w \prod_{t=1}^{T} P(X_t^{(s)}, Y_t | \lambda^{(EV)}, w) \tag{7}
\]

This estimation is conducted by iteratively maximizing the following auxiliary function with EM algorithm:

\[
Q(w; \hat{w}) = \sum_{t=1}^{T} \sum_{m=1}^{M} \gamma_{m,t}^{(s)} \log P(X_t^{(s)}, m | \lambda^{(EV)}, \hat{w}) \tag{8}
\]

where

\[
\gamma_{m,t}^{(s)} = P(m | X_t^{(s)}, \lambda^{(EV)}, \hat{w}). \tag{9}
\]

The ML estimate of the weight vector is given by

\[
\hat{w} = \left\{ \sum_{m=1}^{M} \gamma_{m,t}^{(s)} B_m^{(X)} \right\}^{-1} \sum_{m=1}^{M} B_m^{(X)} \hat{X}_m^{(s)}. \tag{10}
\]

where

\[
\gamma_{m,t}^{(s)} = \sum_{t=1}^{T} \gamma_{m,t}^{(s)}. \tag{11}
\]

\[
\hat{X}_m^{(s)} = \sum_{t=1}^{T} \gamma_{m,t}^{(s)} (X_t^{(s)} - b(0)) \tag{12}
\]

The source independent GMM [5] is used in the first E-step, i.e., the calculation of Eq. (9). And then, the updated canonical EV-GMM is employed in the following steps.

2.4. Conversion with Adapted EV-GMM

The conversion process is straightforwardly performed with the adapted EV-GMM. This paper employs the trajectory-based conversion method considering dynamic features and the global variance (GV) [4].

2.5. Drawback of Conventional Algorithm

Figure 1 shows the schematic image of the adaptation process. Basically the unsupervised adaptation of the EV-GMM is successfully performed using only a small amount of adaptation data such as a few sentences [5] because the speaker dependent characteristics are efficiently modeled on the subspace spanned by the small number of representative vectors. Even if the amount of adaptation data is limited for some mixture components (e.g., the mixture component "G4" in (a) of Figure 1), every mixture component is adapted based on the estimated weight vector. However, if the amount of adaptation data further decreases, the conversion performance rapidly degrades [6] due to the over-fitting problem. In a worst-case, it is possible that each mixture component is assigned to an improper acoustic space as shown in (b) of Figure 1. The resulting adapted GMM often causes the quality and intelligibility degradation of the converted speech due to the feature conversion over different phonemic spaces. This over-fitting problem would be alleviated by reducing the number of estimated weights. However, its reduction process causes another problem, i.e., the degradation of the conversion performance when the amount of adaptation data is enough because the limited number of representative vectors increases the projection error of the mean vectors onto the subspace.

3. MAP Adaptation for Many-to-One EVC

The over-fitting problem seems to be caused by lacks of information on the weight vector to be estimated. We propose the MAP adaptation for the EVC considering prior information on the weight vector for improving robustness of the EV-GMM adaptation against the amount of adaptation data,
adaptation data. If the adaptation data is not given, i.e., zero between them dynamically changes according to the amount of adaptation data. Consequently, it causes the best conversion performance among the many-to-one VC methods when using any amount of adaptation data.

3.1. Training of prior probability density function

We employ the following Gaussian distribution as the prior distribution:

\[ P(w|\lambda^{(w)}) = N(w; \mu^{(w)}, \tau^{-1} \Sigma^{(w)}) . \tag{13} \]

A model parameter set \( \lambda^{(w)} \) consisting of the mean vector \( \mu^{(w)} \) and the covariance matrix \( \Sigma^{(w)} \) is trained in advance using a set of weight vectors estimated for individual pre-stored source speakers in SAT (see Eq. (4)) as follows:

\[ \hat{\lambda}^{(w)} = \arg \max \prod_{s=1}^{S} P(\hat{w}_s|\lambda^{(w)}). \tag{14} \]

Note that the hyper-parameter \( \tau \) is set to 1 during this training.

3.2. MAP adaptation of EV-GMM

For given the adaptation data, the MAP adaptation of the EV-GMM is conducted as follows:

\[ \hat{w} = \arg \max \prod_{t=1}^{T} P(X_t^{(w)}|\lambda^{(w)}) \prod_{t=1}^{T} P(X_t^{(EV)}, w). \tag{15} \]

This estimation is conducted by iteratively maximizing the following auxiliary function with EM algorithm:

\[ Q(w; \hat{w}) = \log P(w|\lambda^{(w)}) + \sum_{t=1}^{T} \beta_{m,t} \log P(X_t^{(w)}, m|\hat{w}). \tag{16} \]

The MAP estimate of the weight vector is given by

\[ \hat{w} = \left\{ \tau \Sigma^{(w)-1} + \sum_{m=1}^{M} \beta_{m,t} \Sigma_m^{(w)-1} \sum_m^{(w)-1} B_m^{(w)} \right\}^{-1} \left\{ \tau \Sigma^{(w)-1} \mu^{(w)} + \sum_{m=1}^{M} B_m^{(w)} \Sigma_m^{(w)-1} X_m \right\} . \tag{17} \]

The MAP estimate is calculated by linear interpolation considering covariance values between the prior mean vector \( \mu^{(w)} \) and the ML estimate shown by Eq. (10). An interpolation rate between them dynamically changes according to the amount of adaptation data. If the adaptation data is not given, i.e., zero occupancy probabilities shown by Eq. (9), the MAP estimate is equal to the prior mean vector. As the amount of adaptation data increases, the MAP estimate asymptotically approaches to the ML estimate. Note that if the hyper-parameter \( \tau \) is set to 0, i.e., ignoring the prior information, the MAP estimate is always equal to the ML estimate.

4. Experimental Evaluations

4.1. Experimental Conditions

We used 160 speakers consisting of 80 male and 80 female speakers in Japanese Newspaper Article Sentences (JNAS) database [10] as the pre-stored source speakers. Each of them uttered a set of phonetically balanced 50 sentences. We used a male speaker not included in JNAS, who uttered the same sentence sets as uttered by the pre-stored source speakers, as the target speaker. We used ten test source speakers consisting of five male and five female speakers, who were not included in the pre-stored source speakers. Those speakers uttered 53 sentences that were also not included in the pre-stored data sets. The number of adaptation sentences was varied from 1/16 (around 300 ms) to 16. We used 21 sentences not included in the adaptation data as the evaluation data. More detail conditions are described in [5, 6]. All speech data were sampled at 16 kHz.

We used mel-cepstrum as a spectral feature. The first 24th mel-cepstral coefficients were extracted from speech data. The STRAIGHT analysis method [11] was employed for the spectral extraction. A simple linear conversion with means and standard deviations of log-scaled \( F_0 \) of the source and the target speakers was employed in the \( F_0 \) conversion.

We evaluated the performance of the proposed MAP adaptation compared with those of the conventional many-to-one VC algorithms including source speaker-independent conversion, speaker selection, and EVC (with SAT), which were evaluated in [6]. The number of mixture components was set to 128 in every conversion algorithm. The number of representative vectors of the EV-GMM was set to 159. The hyper-parameter \( \tau \) was set to a constant value, which was manually determined in our preliminary experiment.

4.2. Objective Evaluations

We conducted the objective evaluation using the mel-cepstral distortion between the converted and target mel-cepstra as an evaluation measure. The averaged distortion over all test speakers was 8.11 [dB] before the conversion.

Figure 2 shows the mel-cepstral distortion in each conversion method when varying the number of adaptation sentences. As reported in [6], EVC is the most effective compared with the speaker independent conversion and the speaker selection when the number of adaptation sentences is more than one. However, its conversion performance starts to degrade due to the over-fitting problem when the number of adaptation sentences decreases less than one. As a reference, results of EVC using only nine representative vectors are also shown in this figure. As mentioned in Section 2.5, although the reduction of the number of adaptation parameters alleviates the over-fitting problem, it causes another problem of the performance degradation when using more than one adaptation sentence. The proposed MAP adaptation remarkably alleviates the over-fitting problem while keeping the conversion performance enough high when using enough amount of adaptation data. Consequently, it causes the best conversion performance among the many-to-one VC methods when using any amount of adaptation data.
can observe that density of adaptation parameters was successful applied to the
that only the proposed methods because the over-fitting problem happens, the
pair-combination of all conversion methods. The number of listeners, and then they were asked which voice sounded more natural. Each listener evaluated 120 stimulus-pairs including every
stimulus-pairs including every pair-combination of all conversion methods. The number of listeners was seven.
Results of the preference tests are shown in Figure 3. We can observe that (a) when the number of adaptation sentences to 1/16, the EVC is considerably worse than the other three methods because the over-fitting problem happens, (b) when the number of adaptation sentences to 1, all methods have comparable performance, (c) when the number of adaptation sentences to 16, the proposed MAP adaptation and the EVC significantly outperform the other two methods. These results demonstrate that only the proposed MAP adaptation is always in the best group when using any amount of adaptation data.

5. Conclusions
This paper described a novel many-to-one voice conversion (VC) algorithm based on eigenvoice conversion (EVC) and maximum a posteriori (MAP) adaptation. The prior probability density of adaptation parameters was successfully applied to the EVC adaptation process for alleviating the over-fitting problem, which is often observed when using the very limited amount of adaptation data. Results of objective and subjective evaluations demonstrated that the proposed method seamlessly works for various amounts of adaptation data while keeping its performance the same or more than the best one among the other many-to-one VC algorithms for each amount of adaptation data.

Acknowledgment: This research was supported in part by MEXT Grant-in-Aid for Young Scientists (A). The authors are grateful to Prof. Hideki Kawahara of Wakayama University, Japan, for permission to use the STRAIGHT analysis-synthesis method.

6. References